

# Máster en Programación avanzada en Python para Big Data, Hacking y Machine Learning

Programación Python para Machine Learning

# LECCIÓN 04

## Selección de características

# ÍNDICE

- ✓ Introducción
- ✓ Objetivos
- ✓ Principios de la selección de características
- ✓ Métodos Filter de selección de características
- ✓ Métodos Wrapper de selección de características
- ✓ Análisis de Componentes Principales
- ✓ Conclusiones

# OBJETIVOS

Al finalizar esta lección serás capaz de:

1

Entender el concepto de selección de características y las razones para su aplicación.

2

Conocer los distintos tipos de técnicas para selección de características.

3

Aplicar métodos Filter y Wrapper para selección de características.

4

Conocer los conceptos del PCA y su aplicación.

# INTRODUCCIÓN

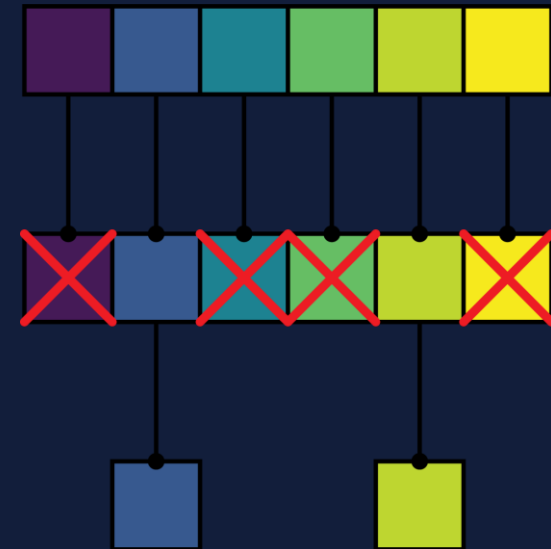
- Tamaño de los conjuntos procedentes de problemas en el mundo real.
- Dificultad de procesamiento.
- Necesidad extra de recursos
- Atributos que son irrelevantes, redundantes o directamente que introducen ruido.



## SELECCIÓN DE CARACTERÍSTICAS

Proceso de selección de las características más significativas o relevantes de un conjunto de datos determinado.

- | Se reduce la complejidad de un modelo y facilita su interpretación.
- | Permite que el entrenamiento del modelo de sea más rápido.
- | Mejora el rendimiento de los modelos.
- | Reduce el sobreentrenamiento (overfitting)



## MÉTODOS FILTER PARA SELECCIÓN DE CARACTERÍSTICAS

Grupo de técnicas que tratan la elección de un subconjunto de atributos:

1. sin incluir ningún algoritmo de Machine Learning.
2. solo basándose en las propiedades estadísticas de los datos.

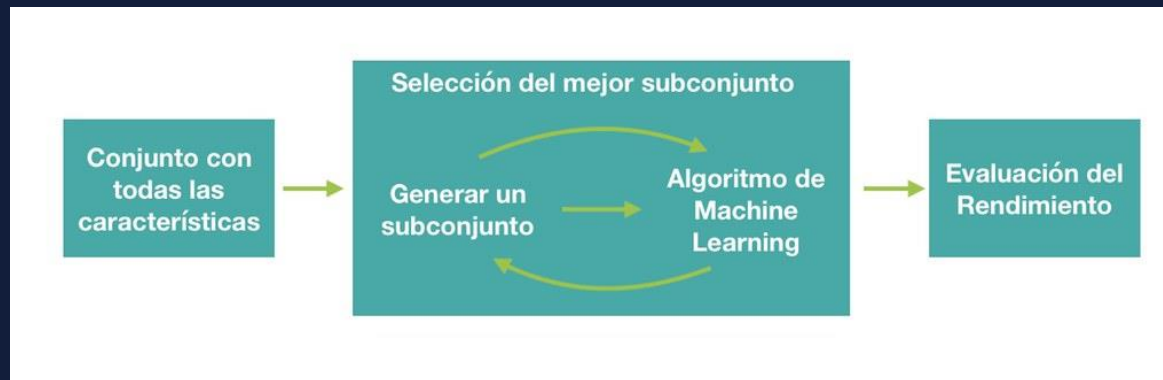
Las características se clasifican utilizando estadísticos que tienden a determinar la correlación de las características con la variable de objetivo.

Variables de entada	Regresión	Clasificación
Numéricas	Pearson, Spearman	Anova, Kendall, LDA
Categóricas	Anova, Kendall	Chi-cuadrado, MI

## MÉTODOS WRAPPER PARA SELECCIÓN DE CARACTERÍSTICAS

Grupo de técnicas que necesita de un algoritmo de Machine Learning, utilizando su rendimiento como criterio de evaluación del subconjunto de atributos seleccionados.

Naturaleza iterativa o recursiva.





## CONSIDERACIONES PRÁCTICAS

- ✓ Métodos Filter más rápidos que los Wrapper.
- ✓ Los métodos Filter utilizan estadístico, los Wrapper validación cruzada.
- ✓ Los métodos Wrapper se acercan más al subconjunto óptimo.
- ✓ Overfitting más probable con métodos Wrapper.

¿En qué punto del proceso ML hay que proceder con la FS?

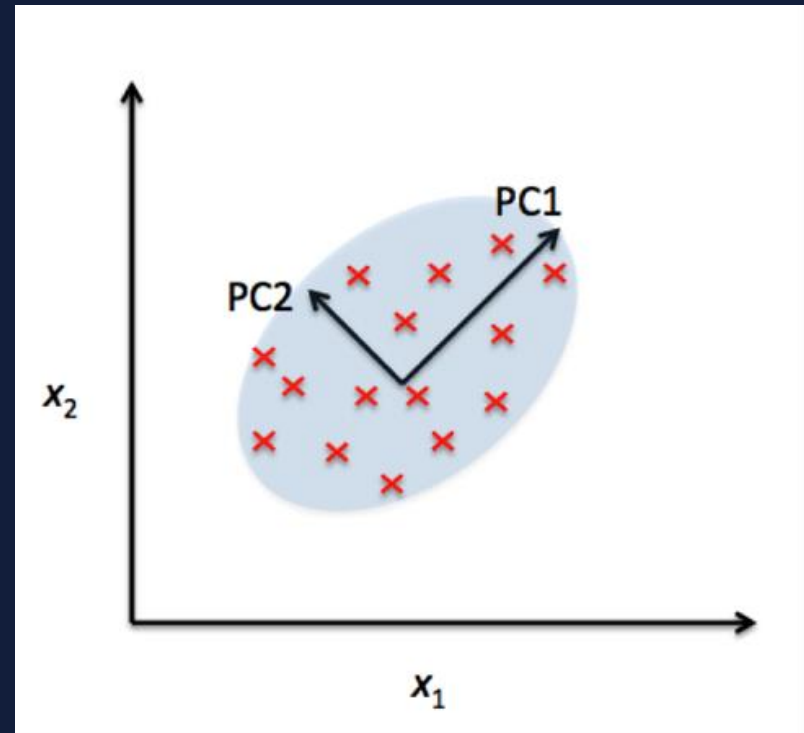
## PCA: Análisis de Componentes Principales

Reducción de la dimensionalidad.

No es selección de características.

Describe el conjunto de datos en nuevas variables (componentes) combinación lineal de las originales ortogonales entre sí.

Muy utilizado para análisis exploratorio.



MUCHAS GRACIAS POR SU ATENCIÓN



[jperez@grupomainjobs.com](mailto:jperez@grupomainjobs.com)



Javier Pérez Rodríguez  
[www.linkedin.com/in/perezxavi](http://www.linkedin.com/in/perezxavi)



[twitter.com/eiposgrados](https://twitter.com/eiposgrados)



[facebook.com/eiposgrados](https://facebook.com/eiposgrados)



[instagram.com/eiposgrados](https://instagram.com/eiposgrados)